When Stale Facts Win: Why Language Models Lose Context in Corrections, and How to Measure, Predict, and Fix It

Anonymous authors
Paper under double-blind review

Abstract

If a context mentions a fact and later explicitly corrects it, a language model should respond with the new value. Often the old value remains in effect. We investigate this failure as a distinct ability to track updates and distinguish it from the mere retrieval of information. A causal audit first examines, for each model, whether attention-based sentence rankings are in fact necessary and sufficient for the response. A pattern emerges: the models focus on the correction but still select the outdated value. Using StaleBench, we measure this behavior across different context lengths, update types, and model sizes. The results show that update tracking neither reliably scales with model size nor is predicted by standard long-context tests. A position swap also provides evidence that smaller models often follow a positional heuristic; for large models, this finding regarding the mechanism remains tentative. Common prompt and decoding methods do not robustly resolve the problem. We therefore evaluate a training-free context pruning method that removes outdated value-binding sentences before generation. It significantly improves the responses, provided that the conflicting statements are detected. It is this detection that proves to be a central bottleneck on natural Wikipedia text. This paper does not offer the familiar recommendation to delete old values, but rather a causally verified measurement and diagnostic procedure, a benchmark for update tracking, and an explicit analysis of the assumptions and limitations of context-based repairs. In practice, this implies that agent stores and retrieval systems should resolve conflicting versions prior to generation and test update tracking separately from information retrieval.

1 Introduction

Agents based on large language models operate on long, evolving contexts, such as conversation histories and tool outputs. Common architectural proposals for agent memory keep such outputs in-context and update them (Packer et al., 2023), while retrieval systems assemble contexts from individually retrieved passages (Lewis et al., 2020). In both cases, facts are overwritten. For example, a user corrects a number, a tool provides a newer measurement, or a note replaces an older one. The implicit assumption behind this is that the model operates using the most recent valid value. This work systematically demonstrates that this assumption is incorrect for many models, explains why it is incorrect, when it is incorrect, and what can be done to address it.

Assume a document lists a catalog number, explicitly corrects it later, and the model is then asked for the current number. Many models respond with the old number. We refer to this behavior as a "stale answer." The novelty of this work lies not in the well-known recommendation to delete old values, but in the causally verified measurement of the error, its distinction from ordinary information retrieval, and the systematic examination of the assumptions under which a context-based correction is effective. This paper brings together four interrelated contributions:

1. Before using attention protocols as evidence of what a model has read, we check for each model whether the sentences receiving the most attention are causally necessary (masking disrupts the answer) and sufficient (they alone account for the answer), using random and zero-context controls. Nine out of ten models pass; Phi-4-mini generates answers from

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

duplicates that it does not visibly attend to. Reliability must be verified for each model; it cannot be assumed.

2. StaleBench, a benchmark for tracking updates. 1,000 deterministically generated examples across four context lengths (1k to 16k tokens), five overwriting mechanisms, one to four overwrites, and distractors. Seven models ranging from 0.5B to 70B reveal three structural findings: First, performance does not increase monotonically with model size. Second, the length axis is inverted; short contexts are more difficult than long ones. Third, established long-context tests (Needle In A Haystack (NIAH), variable tracking) do not predict failure.
3. A positional swap distinguishes between two possible strategies: following the position of the last value or the semantic marker of an update. Small models predominantly follow the position; a single 70B model more frequently follows the semantics. We treat this contrast as a hypothesis about different strategies, not as a proven scaling effect. Among the countermeasures, neither prompt instructions nor chain-of-thought nor context-aware decoding are reliable across models. In contrast, training-free context cleaning provides robust help once the conflicting sentences are identified.
4. Validation on real text. In Wikipedia articles containing factual statements taken verbatim, the problem occurs even outside the templates for the small models tested. At the same time, this track highlights the critical limitation of the repair process: when dealing with natural text, the bottleneck is not the removal but the reliable identification of all value-laden statements. The real-text track thus evaluates transfer but does not allow conclusions to be drawn about frontier models.

Practical implications: Agent memories that append corrections rather than deleting old values leave the old value in-context, where it often prevails and destabilizes even correct answers. Retrieval pipelines in the style of Lewis et al. (2020), which assemble isolated snippets, create the conditions under which the old value most frequently wins in our measurements.

2 Related Work

Jain & Wallace (2019) showed that attention weights can often be replaced without changing the prediction; Wiegreffe & Pinter (2019) and Serrano & Smith (2019) clarified the debate. We do not settle the question in general terms, but rather provide a test for each model and task to determine whether attention-based rankings are causally necessary and sufficient. Under identical conditions, both outcomes occur. Causal interventions of this kind are also used in the model editing literature (Meng et al., 2022).

Previous studies have primarily examined conflicts between parametric knowledge and context (Longpre et al., 2021; Xie et al., 2024) or between multiple source documents (Hou et al., 2024). Our focus is on conflict *within* a context, an update within the same document or conversation; we filter out parametric effects using a zero-context control.

NIAH (Kamradt, 2023), RULER (Hsieh et al., 2024), and BABILong (Kuratov et al., 2024) measure whether information is found in a long context. Our results show that retrieval and updating are distinct abilities: models with perfect NIAH performance fail at update tracking within the same context window (Section 4.3). Position effects such as "lost-in-the-middle" (Liu et al., 2024) and "attention sinks" (Xiao et al., 2024) affect the process of finding values; in our approach, attribution is perfect in all contradiction runs. What fails is the choice between two found values.

Model editing (Meng et al., 2022; Zhong et al., 2023) modifies weights, while in-context editing (Zheng et al., 2023) and context-aware decoding (Shi et al., 2024) modify the prompt or decoding. We evaluate instruction- and decoding-based methods as baselines and compare them to a masking-based repair method that requires no training and no access to weights beyond logits and masks.

To the best of our knowledge, none of the studies mentioned accomplishes four things. First, the attribution method is causally verified *before* it is used as evidence for behavioral claims, with the result that verification fails for a common model. Second, the conflict is strictly isolated *within* a context: the zero-context control filters out parametric effects, which are the subject of the knowledge-conflict literature. Third, the findings regarding mechanisms are intervention-based rather than correlational: the attractor effect and position swapping are based on controlled mask and position

manipulations with preserved position IDs. Fourth, a training-free repair is subjected to the same conditions as the instruction and decoding baselines, including an adversarial test that challenges its own core assumption.

3 The Causal Audit: Method and Results

Each document interweaves eight predefined facts ("The catalog number for product *Vqxrtz* is 4831.") with filler sentences drawn from a fixed pool, about 49 sentences per document. Since the facts are predefined, their exact positions in the text are known down to the character. Three modes: `plain` (each fact appears once), `duplicate` (the queried fact is repeated verbatim), `contradiction` (the queried fact first appears with an incorrect value and is later corrected; the final value is the ground truth). The generator creates two surface formulations from identical random seeds: one security-related (vault codes) and one neutral (product catalog numbers); entities, values, and positions are identical.

We decode using a greedy approach with eager attention and, for each generated token, accumulate the attention weight of the last query position for each document sentence, averaged across heads and the lower half of the layers. Four interventions set entries in the attention mask corresponding to prompt positions to zero, with explicitly provided position IDs, so that gaps do not shift positions: *necessity* (hide the top-k sentences; the answer must break), *random check* (hide k other sentences; the answer must survive), *sufficiency* (hide everything except the top-k; the answer must survive), and *zero-context* (hide everything; filters out memorized knowledge). Each condition is evaluated twice: via string matching and via the teacher-forced log-likelihood of the correct answer. Safety-tuned models sometimes withhold the answer under strong masking; we classify each output as correct, withheld, or incorrect, and consider a condition to preserve evidence if its log-likelihood drop remains below half of the document's own necessity drop. The intervention mask is generated in the harness before the prefill and applied across all layers both during the prefill and at every autoregressive decoding step; masked sentences therefore cannot preemptively write their information into other residual streams. Averaging over the lower half of the layers is a predefined ranking heuristic, not a claim of its superiority; the causal audit therefore checks the resulting rankings separately for each model.

Across ten checkpoints (Qwen2.5 0.5B to 14B, Base and Instruct; Llama-3.2-3B, Llama-3.1-8B; Phi-4-mini), Precision@3 in plain and duplicate modes is 100 percent for nine models; masking the top 3 breaks the answer in every usable document (mean log-likelihood drops of $+2.4$ to $+12.9$ nats), random masking makes no difference (0.00 nats), and the top 3 alone are sufficient 100 percent of the time under a refusal-aware evaluation. Phi-4-mini fails: in duplicate mode, masking its top 3 breaks the answer in only 20 to 40 percent of the documents; in all 42 unbroken documents, the top 3 covered only one of the two copies, and the model responded based on the other. Its attribution also depends on the phrasing (P@3: 97 percent for the vault phrasing, 66 percent for the neutral phrasing). Phi-4-mini's task performance is unremarkable; what breaks down is the connection between attending to and using the information.

In every contradiction run, the corrected sentence is the highest-ranked attribution, and yet the models respond with the outdated value in 7 to 93 percent of the documents (Table 1). Size does not resolve this issue within the measured range: the Qwen2.5-Instruct scale yields the following results for the neutral formulation: 73/63/63/27/87 percent for 0.5B/1.5B/3B/7B/14B; the largest model performs the worst, and the central contrast between 14B and 7B is, despite $n = 30$, significant (87 percent, Wilson interval 70 to 95, versus 27 percent, 14 to 44; the intervals do not overlap). Nor is this finding specific to a single family: within Llama, the larger 3.1-8B (93 percent, 79 to 98, vault) exceeds the smaller 3.2-3B (77 percent, 59 to 88). The gap between 7B and 14B already exists between the base checkpoints (13 versus 57 percent), so it is visible even before instruction tuning; instruction tuning further increases the rate in three out of four measurable cells (7B neutral $13 \rightarrow 27$, 14B neutral $57 \rightarrow 87$, 14B vault $37 \rightarrow 47$).

The old value is an attractor. Two interventions reveal the mechanism. First, hiding three random, irrelevant sentences causes correct answers to become incorrect: in Qwen2.5-3B, this occurred in 7 out of 20 documents answered correctly, with all 7 answers being exactly equal to the outdated value, while the log-likelihood of the correct answer remained virtually unchanged ($+0.2$ to $+0.4$

Table 1: Audit, contradiction mode: percentage of documents answered with the outdated value ($n = 30$ per cell; Wilson 95% intervals for this n are about 15 points wide; the cells that are contrasted in the text do not overlap). The attribution is perfect in all rows except Phi-4-mini: the models read the correction but still respond with the old value.

Model	**Variant**	**Outdated answers (% of 30)**	
		vault	neutral
Qwen2.5-0.5B	Instruct	43	73
Qwen2.5-1.5B	Instruct	70	63
Qwen2.5-3B	Instruct	30	63
Qwen2.5-7B	Instruct	**7**	**27**
Qwen2.5-14B	Instruct	47	**87**
Qwen2.5-7B	Base	13	13
Qwen2.5-14B	Base	37	57
Llama-3.2-3B	Instruct	77	73
Llama-3.1-8B	Instruct	**93**	77
Phi-4-mini	Instruct	50	43

nats vs. $+5.7$ in the case of true evidence removal). The mask does not remove any information; it tips the balance in a close decision between two present values. Second, partial context is worse than either extreme: if only the sentence with the highest attention score is retained (always the correction), both 3B models achieve 100 percent; starting at $k = 3$ the retained set also includes the outdated sentence, causing the accuracy to plummet (Qwen: 55 percent at $k = 3$, 0 percent at $k = 5$ and $k = 8$), even though the same documents are answered correctly when the full document is given. Retrieval pipelines that assemble isolated snippets produce exactly this intermediate condition.

Llama-3.2-3B refuses to provide answers for heavily masked vault-code documents; all 38 sufficiency errors under string evaluation are refusals, not incorrect answers. The log-likelihood criterion evaluates the evidence in 100 percent of these documents as sufficient, and the neutral reformulation completely eliminates refusals (0 out of 180 documents). Masking studies on safety-tuned models must treat refusal as a separate output category; otherwise, safety tuning appears disguised as an interpretability finding.

4 StaleBench: A Benchmark for Update Tracking

The audit measures, but on a small scale. To evaluate update tracking as a capability, we scale up to a benchmark, StaleBench, with 1,000 deterministically generated examples across four axes: *context length* (levels L1, L4, L8, L16 with approximately 1k, 4k, 8k, and 16k tokens), *overwriting mechanism* (five forms: marked update "was changed to," unmarked reassertion, negation "is no longer x, but y," referential update "the value mentioned at the beginning," dialog form), *number of overwrites* (1, 2, or 4), and *distractors* (0 or 3 similar entities with their own values). The question always asks for the last valid value; each answer is additionally assigned to the version that reproduces it. In addition to accuracy, we log attribution precision, necessity, and sufficiency as in the audit; long contexts are processed by a chunked prefill without attention tracking, which materializes attention rows only for decoding steps and is semantically identical to the short-context path. The main comparisons relate to predefined populations and are reported using Wilson intervals. We consider finely resolved combinations of individual factors to be descriptive due to their smaller cell sizes; the results from the large models come from the same stratified subset and are not interpreted as population-wide rankings.

Table 2 shows the main results; three findings stand out.

Table 2: StaleBench main results. Five local models on all 1,000 examples; the two large models (†) on a stratified subset of 100 examples. The length axis is inverted: for the Qwen family, L1 is the hardest and L16 the easiest. In this subset, Qwen2.5-32B outperforms Llama-3.3-70B; Phi-4-mini completely collapses starting at L8.

		Correct by length (%)				
Model	**Overall (%, 95% CI)**	L1	L4	L8	L16	**P@3 (%)**
Qwen2.5-0.5B-Inst.	31 [28–34]	15	23	38	44	72
Qwen2.5-1.5B-Inst.	20 [17–22]	16	15	19	28	75
Qwen2.5-3B-Inst.	48 [45–51]	28	42	50	70	81
Llama-3.2-3B-Inst.	49 [46–52]	45	54	45	51	88
Phi-4-mini-Inst.	25 [22–27]	60	43	**0**	**0**	40
Llama-3.3-70B-Inst.†	91 [84–95]	75	90	100	100	94
Qwen2.5-32B-Inst.†	**99** [95–100]	100	100	95	100	100

4.1 Size does not scale monotonically

Qwen2.5-1.5B (20 percent, Wilson interval 17 to 22) is below Qwen2.5-0.5B (31 percent, 28 to 34); in the observed stratified sample of 100, Qwen2.5-32B (99 percent, 95 to 100) is also above Llama-3.3-70B (91 percent, 84 to 95). Together with the 14B anomaly from the audit, these data contradict a simple monotonic scaling with the number of parameters. However, family, training, and size are confounded in our model set; we do not derive a ranking of model classes from this. We expressly do not claim that frontier models fail. At the upper end, the synthetic benchmark reaches saturation, while the remaining errors of the 70B model are concentrated on short contexts and unmarked reassertions. The error class remains practically relevant for small and medium-sized models, which are deployed locally or at the edge for cost, latency, and privacy reasons and also fail significantly in the real-text track.

4.2 The length axis is inverted

Contrary to what one might expect based on long-context intuition, the accuracy of Qwen models increases with context length (0.5B: 15% to 44%; 3B: 28% to 70%; and 70B: 75% to 100%). In short contexts, the old and new values are close together and compete directly; distance mitigates this conflict. Accordingly, distractors have little negative impact and sometimes even help (Llama-3.2: 42 vs. 56 percent), which is plausible because additional update sentences reinforce the "pay attention to changes" pattern. The strongest factor is the number of overwrites: for the small models, accuracy drops from an average of 54 percent with one overwrite to 20 percent with four overwrites; the most disruptive mechanism is the unmarked reassertion, while the least disruptive is explicit negation.

Phi-4-mini drops from L8 onward to exactly 0 percent, with attribution accuracy also plummeting. Section 4.3 demonstrates that this is not an update issue: NIAH also reports 0 percent for Phi-4-mini starting at L8. The model generally does not process contexts of this length in our setup.

4.3 Retrieval tests do not predict update tracking

Table 3 presents three tasks in identical contexts: classic NIAH (a fact mentioned only once is queried), variable tracking via alias chains (RULER-style), and our update tracking. All functional models achieve 100 percent NIAH while simultaneously achieving only 62 to 92 percent update tracking. A perfect NIAH score says nothing about behavior when data is overwritten. Finding and updating are separate capabilities and must be measured separately.

Table 3: Three tasks in the same context window (% correct, $n = 24$ per cell from L1 to L16, 6 per level). NIAH is perfect everywhere, but update tracking is not; the masking repair closes the gap. *Phi-4-mini, including the L8/L16 collapse, which affects all tasks equally (L1 only: NIAH 100, update 83).

Model	NIAH	Variable tracking	Update tracking	with repair
Qwen2.5-3B-Inst.	100	79	67	100
Llama-3.2-3B-Inst.	100	58	62	100
Phi-4-mini-Inst.	33*	11*	28*	33*
Llama-3.3-70B-Inst.	100	100	92	100

Table 4: Position swap (%; $n = 96$ per local model with Wilson 95% intervals; †70B: $n = 6$, cloud run). After the swap, most small models follow the position (they return the original value mentioned last), while the tested 70B model more frequently follows the semantic update marker.

Model	normal, correct	swapped, correct	follows pure position
Qwen2.5-0.5B-Inst.	31 [23–41]	41 [31–51]	55
Qwen2.5-1.5B-Inst.	38 [28–47]	41 [31–51]	59
Qwen2.5-3B-Inst.	55 [45–65]	30 [22–40]	70
Llama-3.2-3B-Inst.	75 [65–83]	38 [28–47]	60
Phi-4-mini-Inst.	85 [77–91]	34 [26–44]	59
Llama-3.3-70B-Inst.†	100	**83**	**17**

5 Mechanism and Countermeasures

5.1 Position swap: position or semantics?

Does a model respond with the last value because it appears last, or because it is marked as an update? We distinguish between the two by swapping the positions of the original statement and the update statement in single-update examples: the semantically marked update ("was changed to") now appears *before* the original statement. A model that follows the marker remains correct; a model that follows the position now responds with the original value.

Table 4 provides intervention-based evidence of different strategies: after the swap, the majority of the small models follow the position, while the tested 70B model more frequently sticks with the semantically marked update. Because the finding at the upper end is based on a single model and model size is confounded with model family, we do not interpret this as a proven scale transition, but rather as a hypothesis. Position and semantic strategies normally produce the same surface performance and become distinguishable only in the counterfactual swap. Whether the contrast is explained by size, training data, or family must be tested with additional large models.

5.2 Countermeasures

Table 5 compares four countermeasures with the untreated condition: an explicit recency instruction ("Always use the most recently mentioned value in case of contradictions"), chain-of-thought, CAD (Shi et al., 2024), and our repair. Four findings:

Instructions are unreliable. The same instruction moves Qwen2.5-3B by $+33$ points, Llama-3.2-3B by -10, and Qwen2.5-0.5B by 0. This is not an artifact of a single phrasing: with three independently phrased recency instructions, the effects range from -27 to $+27$ points; the same phrasing is the best for one model and the worst for another, and no single phrasing dominates across all models (Appendix A). Anyone who tries to fix their system's stale rate using the system prompt will get a result ranging from a significant improvement to a deterioration.

Table 5: Countermeasures (% correct; L1 and L4, $n = 40$ per model and condition, 70B: $n = 20$). Instructions seem random, CoT yields mixed results, and context-aware decoding (CAD) never helps. With controlled conflict detection, masking repair is the only condition that achieves 90 to 100 percent across all models.

Model	plain	Instruction	CoT	CAD	Repair
Qwen2.5-0.5B-Inst.	28	28	21	28	**95**
Qwen2.5-1.5B-Inst.	15	30	42	15	**90**
Qwen2.5-3B-Inst.	32	65	70	22	**95**
Llama-3.2-3B-Inst.	58	48	75	55	**100**
Phi-4-mini-Inst.	52	60	45	50	**95**
Llama-3.3-70B-Inst.	95	100	100	90	**100**
Qwen2.5-32B-Inst.	100	100	98	98	**100**

Chain-of-thought is stronger but inconsistent. It significantly improves several small models, worsens others, and offers little additional benefit for the strongest models. Even step-by-step listing and selection is not a reliable fix across all models.

CAD never helps. In none of the seven cells does CAD outperform the untreated condition. This is consistent with the diagnosis: CAD reinforces the context at the expense of parametric knowledge, but our conflict lies within the context itself; both values are reinforced.

The repair is robust when the conflict is detected. Under controlled synthetic detection, detection and masking achieve high accuracy across all seven models, including Phi-4-mini, whose attention mechanism is unreliable. This isolates the effectiveness of context pruning; it does not prove that real conflicts are reliably detected.

The repair process operates in three steps without training and without relying on weights beyond logits and the attention mask.

Detection. A deliberately simple keyword scan searches for sentences that associate a value with the queried entity; different values indicate a conflict. On the synthetic audit corpus with known sentence templates, it finds all expected conflicts without false positives. This metric is an internal control measure and does not serve as evidence of general detection performance.

Alert. The system reports the candidates, the model's selection, and the log-likelihood difference. Model confidence is not suitable as a warning signal because outdated answers tend to be more confident on average; in contrast, a positional signal perfectly distinguishes the answers on the controlled corpus.

Masking. All value-binding sentences identified as outdated are hidden, position IDs are retained, and the answer is regenerated. In this way, the repair eliminates all previously outdated responses in the audit corpus and achieves high accuracy even on StaleBench. The evaluation thus addresses two questions: what effect does context cleaning have on known conflicts, and how can such conflicts be identified in natural text? The second question remains an open subproblem.

5.3 Circularity test: a backreference after the last update

The benchmark defines the last valid value as correct, and the naive repair retains the last value-binding statement. One might object that the repair thus incorporates into the pipeline precisely the heuristic specified by the problem definition. We test this with an adversarial set that violates the core assumption: behind the last update, there is an additional reference to the original value, which does not make it valid again ("Historical note: the catalog number for product X was v_0 at the start."). The last value-binding statement is now *no longer* the correction. Table 6 shows the results based on 288 examples per model (L1 and L4, marked mechanisms). The naive "last sentence" repair falls to 0 percent and is entirely predictable: it masks the correction, leaves the backreference in place, and returns the original value in virtually every document (287 out of 288). A marker-sensitive variant, which instead retains the last sentence with an explicit update marker, achieves

Table 6: Circularity test (% correct, $n = 288$ per model, Wilson 95% intervals): after the last update, there is a reference back to the original value. The naive "last sentence" repair fails completely by design; the marker-sensitive variant performs robustly. The unrepaired accuracy also decreases because the old value mentioned later acts as an attractor.

Model	plain	naive repair	marker-sensitive repair
Qwen2.5-0.5B-Inst.	3 [2–6]	0 [0–1]	94 [91–96]
Qwen2.5-1.5B-Inst.	29 [24–34]	0 [0–1]	93 [90–95]
Qwen2.5-3B-Inst.	55 [49–61]	0 [0–1]	90 [86–93]
Llama-3.2-3B-Inst.	66 [60–71]	0 [0–1]	93 [90–95]
Phi-4-mini-Inst.	49 [44–55]	0 [0–1]	96 [93–98]

Table 7: Wikipedia track (%, $n = 103$; Wilson 95% intervals). The error rates on real text fall within the range of the synthetic measurements. The repair helps significantly, but its performance declines compared to synthetic text because keyword detection on natural text becomes incomplete.

Model	correct (new)	outdated (old)	other	with repair
Qwen2.5-0.5B-Inst.	24 [17–33]	54 [45–64]	21	69
Qwen2.5-1.5B-Inst.	45 [35–54]	43 [34–52]	13	78
Qwen2.5-3B-Inst.	50 [41–60]	40 [31–49]	10	79
Llama-3.2-3B-Inst.	33 [25–43]	**50** [40–59]	17	75
Phi-4-mini-Inst.	42 [33–51]	47 [37–56]	12	80

90 to 96 percent across all five models. A secondary finding further exacerbates the situation: even the mere backreference drastically reduces the *unrepaired* accuracy (Qwen2.5-0.5B from 28 to 3 percent; it returns the original value in 89 percent of the documents). An old value acts as a stronger attractor the later it is mentioned again, consistent with the positional heuristic from Section 5.1.

The repair is not a tautology of the task definition, but its naive form relies on a positional assumption that does not hold for backreferences; the marker-sensitive form replaces the positional assumption with a semantic one and is, in principle, not applicable to unmarked reassertions. Both are now explicitly part of the procedure.

5.4 Text-level variant

Attention-mask repair requires access to the model mask; closed APIs do not provide this access. The same operation can be performed at the text level: sentences identified as obsolete are deleted before the prompt is constructed, rather than being hidden in the mask. In our comparison, both variants almost always yield the same results in terms of response and accuracy. The text-based variant is therefore technically compatible with any API that accepts text; model access beyond the generation stage is only necessary for the causal audit. However, this consistency merely demonstrates the interchangeability of the intervention, not the quality of the upstream conflict detection.

6 Validation Using Real Text

Synthetic corpora allow for exact ground truth but raise the suspicion that the phenomenon is an artifact of the source sentences. We therefore construct examples from real Wikipedia articles: using phrase search, we find articles that contain population or founding information in sentences that can be quoted verbatim. The original sentence remains verbatim in the article; later in the text, a realistically phrased update with a modified value is inserted, which unambiguously resolves the question of the current value. This track is evaluated on the five smaller local models. It tests the transfer from the templates and detection in natural text, not the performance of frontier models.

Table 7 shows that, for the small models tested, the proportion of outdated responses in natural text is in the same range as in the synthetic corpus. For this class of models, therefore, the phenomenon

is not merely a template artifact; this finding cannot be generalized to large or proprietary models. At the same time, the repair performance drops noticeably because detection becomes the bottleneck on natural text: real articles mention the same value multiple times and in phrasings that a keyword scan misses. Even masking just one source sentence does not correct every outdated answer, because further mentions of the same old value may remain in the article. The real-text track thus confirms the performance issue for the tested models and simultaneously demonstrates why detection is the central challenge that remains to be addressed.

7 DISCUSSION

The four parts come together to form a practical diagnosis. The models tested can find information in a long context and yet still make the wrong choice between an old and a corrected version; standard retrieval tests do not capture this ability. In small models, the position swap suggests a fragile positional heuristic. The single result from the 70B model points to greater use of semantic update markers but is insufficient to draw a general conclusion. Because the old value can act as an attractor and model confidence does not reliably indicate the error, systems should resolve conflicts before generation. The recommendation to delete old values rather than simply appending new ones is not new; what is new is the causal analysis, the benchmark, the separation of retrieval and update operations, and the explicit measurement of the conditions under which context cleanup succeeds or fails.

Our recommendations: First, agent memory should not append corrections, but rather delete or overwrite old values. Second, retrieval systems should not assemble isolated snippets with conflicting values; a correction plus an old value without context is the measurably worst form of context. Third, long-context retention should be tested using update tasks, not just retrieval tasks. Fourth, attention protocols only constitute evidence after a causal audit, and refusals should be included as a separate category in every masking study.

Limitations. The benchmark facts are numerical. This allows for precise answer mapping but covers only one type of fact. Fine-grained combinations of benchmark factors have small cell counts; we therefore treat such slice comparisons as descriptive and base our main conclusions on predefined sets with Wilson intervals. For cost reasons, the large models were run only on a stratified subset. At the upper end, the synthetic benchmark reaches saturation, and the Wikipedia track consists exclusively of small local models. Conclusions regarding proprietary frontier and reasoning models therefore remain open, even though the behavior-based test and the text variant are technically transferable to them. The mechanism findings at the upper end are based on a single 70B model; model size and model family are confounded. Furthermore, averaging over the lower half of the layers is a non-ablated heuristic. The causal audit examines the resulting rankings but does not replace a comparison with attention rollout or gradient-based methods. For the baselines, we evaluate several recency instructions, chain-of-thought, and CAD, but not self-consistency, two-pass prompts, or LLM-based conflict detectors. Above all, the repair requires that conflicts be detected. The perfect synthetic detector score applies only to known templates; on natural text, detection becomes the bottleneck. Finally, the circularity test shows that every repair variant relies on an explicit positional or semantic assumption and fails when that assumption is violated.

AI USE STATEMENT

In this work, we used generative AI tools for code and language improvements, and synthetic dataset creation. We have not used generative AI tools for other tasks with required disclosure. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All runs are deterministic, using fixed random seeds and greedy decoding; the corpus generator, harness, benchmark manifest, evaluation, and all raw results are stored in the anonymized supplementary materials. The five smaller models run on a single 12-GB GPU, while the larger ones run on an A100/H100; a full run of 1,000 examples takes 6 to 14 hours, depending on the model.

REFERENCES

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. WikiContradict: A benchmark for evaluating LLMs on real-world knowledge conflicts from Wikipedia. *Advances in Neural Information Processing Systems*, 37:109701–109747, 2024.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Greg Kamradt. Needle in a haystack: Pressure testing LLMs. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`, 2023.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, 2021.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, 2019.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, 2024.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4862–4876, 2023.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702, 2023.

A Three Examples of Instruction Wording

The three recency instructions are phrased independently (W1: "Always use the most recently mentioned information"; W2: "Values may have been updated; rely only on the most recent information"; W3: "The document contains corrections; base your answer on the latest revision"). Table 8 shows: no model is robust against the formulation, and no formulation wins across the board. W3 is the best for Qwen2.5-0.5B ($+13$) and, for Llama-3.2-3B, by far the worst (-27).

Table 8: Three recency instruction formulations (% correct, $n = 40$ per cell, with the same examples in each row within the run). Effects range from -27 to $+27$ points; depending on the model, the same wording can be either the best or the worst.

Model	**plain**	**W1**	**W2**	**W3**
Qwen2.5-0.5B-Inst.	12	5	12	25
Qwen2.5-1.5B-Inst.	20	35	35	28
Qwen2.5-3B-Inst.	45	65	72	60
Llama-3.2-3B-Inst.	52	50	60	**25**
Phi-4-mini-Inst.	50	60	68	62

B Partial Updates

An update can also arrive *partially*. We ask three questions about the same context that retrieve the same fact via different paths: the value itself, its last digit, and its checksum. On 1,050 documents using five small Instruct models, up to 47 percent of the documents are linked to different values via different paths under a single correction. The dominant trend is counterintuitive: more often than not, the *direct* answer is outdated and the derived one is current (124 versus 28 path pairs), which is plausible because the direct question matches the surface pattern of the outdated sentence, while the derived one does not. With repeated overwriting (two or four updates, as continuous text or a chat log), the current value almost disappears: the latest version accounts for only 12 percent of the direct answers, while the *first* update accounts for 54 percent. The models apparently register one update and then stop counting. Masking the outdated sentences shifts 68 percent of the outdated derived responses to the corrected value, confirming the causality. In practice, this means: an audit that checks only the direct response certifies systems that simultaneously report the corrected value and perform calculations using the outdated one.